Optimizing SDXL Inference on Qualcomm Cloud AI 100 accelerators with MultiDevice Scheduling and DeepCache Techniques

Abstract

Stable Diffusion XL (SDXL) demands significant compute resources, especially in the UNet denoising component. To optimize SDXL inference on Qualcomm Cloud AI 100 accelerators, we implement a multi-device scheduling approach that distributes work across "master" and "worker" devices. This minimizes model switching overhead and boosts throughput by 48% on Ultra devices.

We also apply DeepCache, a method that leverages temporal redundancy in diffusion models to cache and reuse high-level features. DeepCache reduces redundant calculations, improving SingleStream latency by 1.4x on Ultra and 1.6x on Pro cards.

Combining multi-device scheduling and DeepCache on GIGABYTE R282-Z93 servers with Ultra cards yields a 3.03x speedup in the Offline scenario. Our techniques demonstrate significant potential for accelerating SDXL on Qualcomm Cloud AI hardware, enabling more efficient generation of high-quality images with text-to-image models.